

**STAR
and the
Grand Challenge Project**
(<http://www-rnc.lbl.gov/GC/>)

**D. Olson (in absentia)
STAR Collaboration Meeting
20-25 July 1997**

This presentation is prepared for the STAR Collaboration meeting at BNL, July 1997. It is distributed to the members of STAR since I am not able to attend this meeting.

Outline

- People
- The charge of the HENP GC project
- RHIC Computing
- The Problem
- The Approach
- Benefits

People (currently active)

LBNL

NP D. Olson (PI), G. Odyneec, F. Wang, N. Xu, R. Porter

HEP J. Siegrist (PI), I. Hinchliffe, R. Jacobsen

Computing H. Nordberg, C. Tull, D. Quarrie, W. Johnston,
A. Shoshani, D. Rotem

ANL E. May, D. Malon

BNL B. Gibbard, D. Stampf, D. Morrison

FSU G. Riccardi

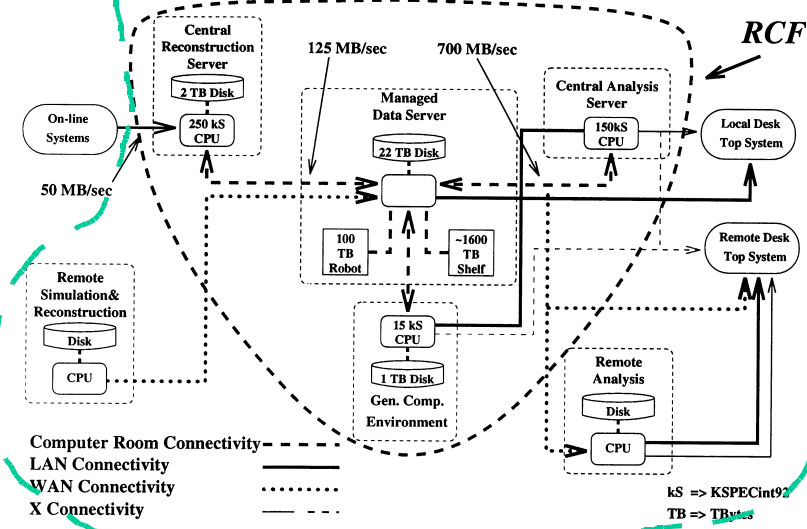
Rice P. Yepes

U.Tenn. S. Sorensen

The Charge

- Demonstrate a solution for data access and analysis for RHIC.
- Three (2.5) year project (FY97, FY98, FY99).

RHIC Computing Model

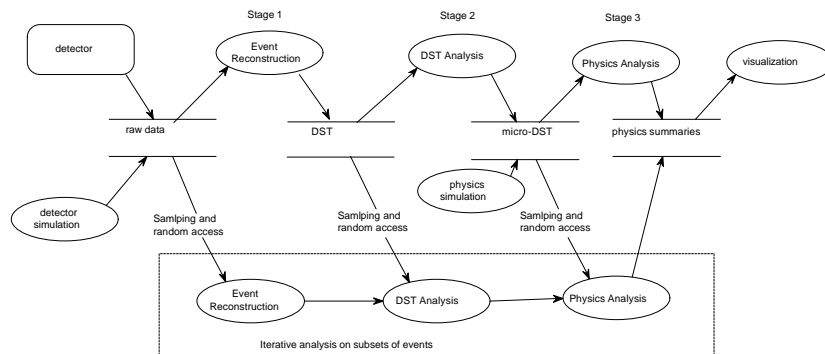


18 July 1997

STAR & the GC, D. Olson

5

Computing Model: Stages/Phases of Processing



18 July 1997

STAR & the GC, D. Olson

6

Stage 1: Event Reconstruction is a single user single pass operation.

Stage 2: DST analysis & micro-DST building is shared across groups of users and occurs a few times for each raw event.

Stage 3: Physics analysis happens at the individual level and occurs several times for each micro-DST, for each group or individual.

Iterative analysis & development: This activities occupies most of the time of physicists doing analysis. It involves accessing small samples of data from all stages (raw through physics summaries). Access of these small samples is repeated many times.

In some cases the small samples should be completely uncorrelated with features in the data except for trigger conditions. This is necessary for tuning algorithms.

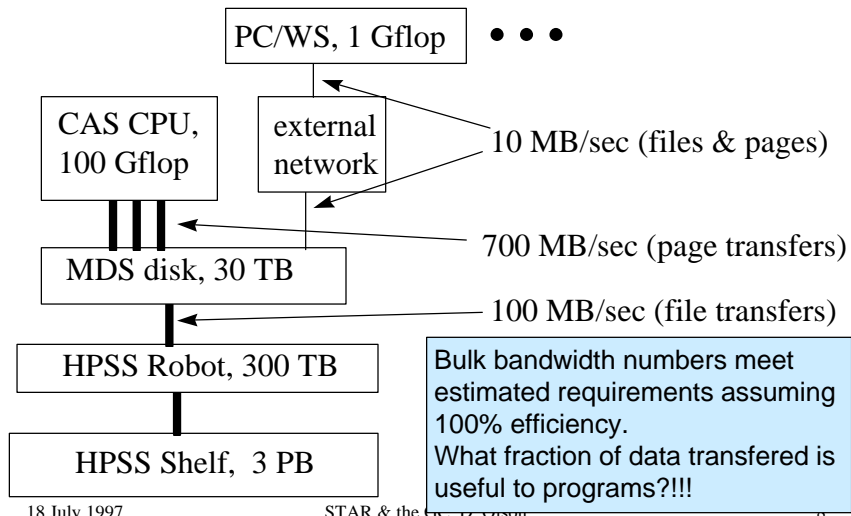
Other times the events are selected based entirely upon features in the data. This is necessary for diagnosing unusual features in the data, either from new physics or (more commonly) from some systematic experimental effects.

Requirements

- Address the tape-disk-cpu data access bottlenecks.
- Data access solution must not preclude requirements spanning RHIC computing
 - event reconstruction (DST production)
 - selections (micro-DST generation)
 - analysis (single process development and PIAF-like parallel processing)
 - simulations (mixing data sources for comparison with theory)
 - robustness (operational efficiency > ??%)
 - tunable system (load balancing for op. efficiency)

The Bottlenecks

(my est. for RHIC capacity, year 3)



Tapes are transfered from shelf to robot.

Files are transfered from robot to disk.

Pages are transfered from disk to CPU memory.

Files on disk are shared by many processes, increasing the useful efficiency of file transfers.

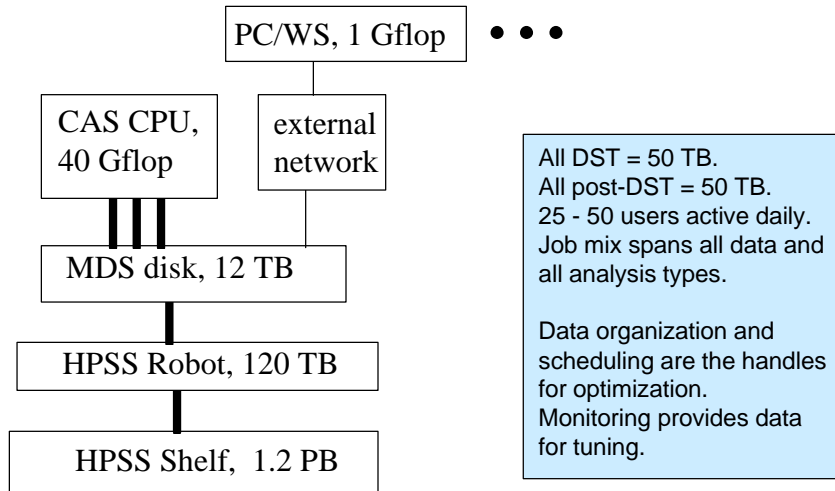
Scheduling analysis tasks which share files can be helpful.

The CPU time per event for analysis tasks in the CAS varies by more than an order of magnitude, which will make scheduling complicated for efficient shared file access.

Pages are transfered from disk to memory on a per process basis. Sharing data across many processes does not increase the efficiency of this transfer.

The Bottlenecks

(my est. for year 3, STAR = 0.4 * RHIC)



18 July 1997

STAR & the GC, D. Olson

9

Analysis activities.

Quick analysis on new data.

Detailed analysis on old data.

In depth studies on unusual effects and (hopefully) new phenomena.

Much iterative development of both analysis algorithms & parameters.

Improvements of calibrations on old data to improve resolutions and reduce systematic effects.

Storage issues.

The sum total of data volume interesting to look at on a daily / weekly basis is much larger than the disk cache.

The handles on efficient tape-disk transfers and disk cache management are:
data organization
scheduling.

The handle on efficient disk-CPU transfers are:
data organization.

Data organization & scheduling

- Define how to order files on tape.
- Define how to map substructures of events onto files (cluster by type).
- Define how order event (substructures) by feature, i.e., trigger streams, filtering, query patterns (cluster by value).
- Coordinate analysis tasks wanting data with the data available on disk.

18 July 1997

STAR & the GC, D. Olson

10

Effective placement of files on tape should result in most files being read sequentially in order to minimize the tape mounting and file seeking times.

Event substructures means defining the contents of raw data, DST, micro-DST, etc. This is organizing the data according to the content type, hence, cluster by type.

In addition to “cluster by type” one can sort data by value of some features, such as trigger type (on-line and off-line), signals uncovered during analysis, variations in analysis algorithms, etc.

Monitoring

- Items to monitor
 - File placement on tape.
 - Fraction of file accessed from disk.
 - Fraction of page used by program.
- Analysis of monitoring data is used to diagnose inefficiencies.
- System should be tunable based on this analysis.

18 July 1997

STAR & the GC, D. Olson

11

File placement monitoring means getting data out of HPSS.

Fraction of file accessed and fraction of page accessed means getting data from the ODMG software (like Objectivity).

Analysis of the monitoring data and subsequently adjusting the system parameters (scheduling, data organization, block sizes) will need to be developed, and should be developed in the context of what can possibly be adjusted.

The Approach

- Adopt an architecture which can address the year 2+ requirements.
- Develop early implementation which can meet year 1- requirements.
- Prototype at NERSC.
- Demonstrate at RCF some possible scenarios with simulated data.

18 July 1997

STAR & the GC, D. Olson

12

The architecture should be capable of satisfying the performance and functionality requirements for the peak load.

The implementation should be a system integration effort of existing software, where possible and only develop new code where necessary.

Candidates for existing software are:

STAF

Objectivity

HPSS

DPSS

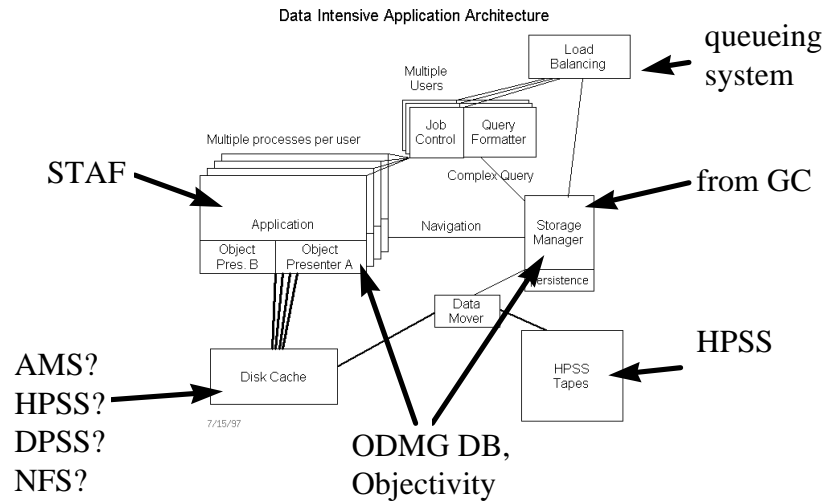
Nile (job control / scheduling for the CLEO experiment)

MPI

Orbix (CORBA)

DQS, Load Leveller (queuing system)

The Architecture (Software)

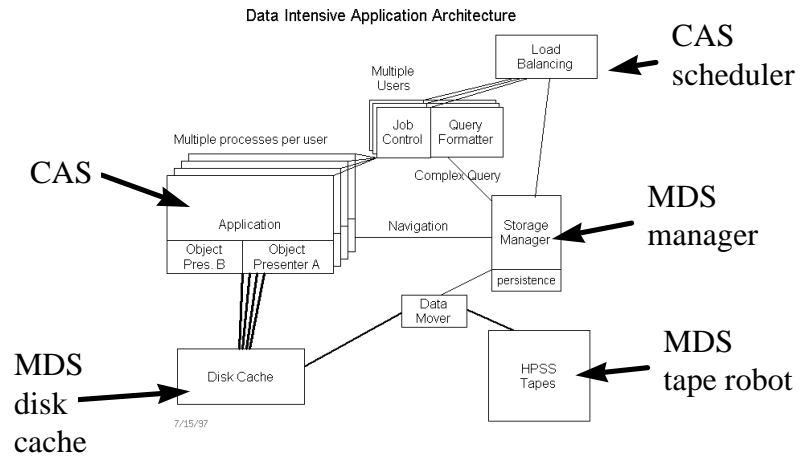


18 July 1997

STAR & the GC, D. Olson

13

The Architecture (Hardware)

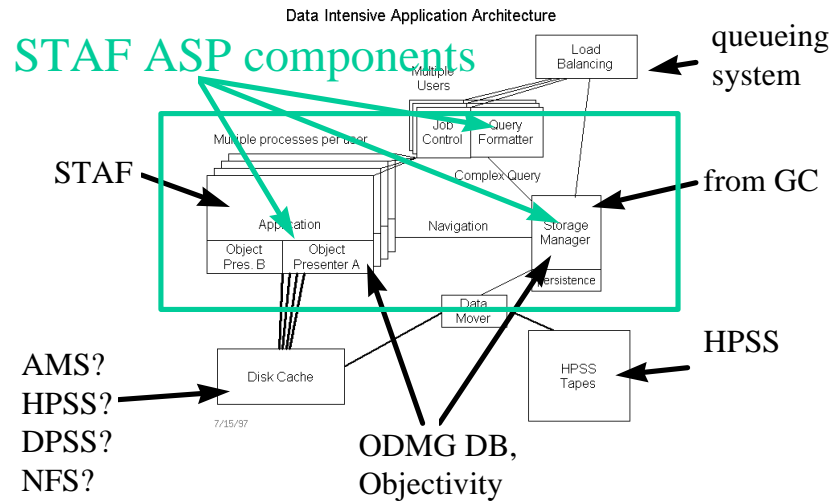


18 July 1997

STAR & the GC, D. Olson

14

Initial Software Prototype



18 July 1997

STAR & the GC, D. Olson

15

Near-term plans

- Develop dataset of simulated events
- Collect data organization ideas from experimental groups (define query/access patterns)
- Investigate HPSS <--> disk issues.
- Investigate ODMG & Objectivity issues.
- Interface STAF to Objectivity.
- Implement prototype of architecture.

18 July 1997

STAR & the GC, D. Olson

16

Simulated dataset:

About 10 TB from event generators & geant. Produce 50 TB datasets for testing with randomizing and embedding from the original 10 TB set.

Data organization:

Discussions of expected logical access patterns as they relate to the various analysis activities are very necessary and useful input.

ODMG & Objectivity:

Determine what scope an Objectivity implementation can have today. Outline a roadmap for full implementation.

Interface STAF to Objectivity:

Develop an ASP (a STAF component) that connects to an Objectivity database. Develop an interface between datasets/tables to ODMG objects.

Prototype implementation:

Develop ASP's for storage manager and query formatter initially. These components can later be distributed via CORBA.

Benefits to STAR

- Develop tools to permit organizing data for efficient access, and re-organization if necessary.
- Analysis system which is tunable.
- Enhance performance as well as “prevent” disastrous behaviour.
- Couples to computing expertise & resources beyond the NP community.

18 July 1997

STAR & the GC, D. Olson

17

Allows the development of tools to handle data. Details can be tested and bugs worked out before we are swamped with real data.

Disastrous behaviour means an analysis task can possibly request data in an order and amount which causes the tape robot to trash around and the disk cache to be used very ineffectively. For example, selecting a completely random sample of events in a random order such that one or a few events is read from every tape will essentially never complete which wasting the tape drive and disk resources. The storage manager in the architecture prevents this behaviour by ordering event access and scheduling tape access.

The people and resources of DOE/MICS contribute to solving the STAR and RHIC computing problem.